N68171-96-C-9110
R&D 8076-EE-01

Optimal Program Scheduling on Parallel and Distributed Systems: Final Report

W F McColl*
Oxford Parallel, OUCL, Wolfson Building,
Parks Road, Oxford OX1 3QD, England

OPSPDS Technical Report 7-97
(October 1997)

1 Introduction

The Department of Defense High Performance Computing Modernization Program (HPCMP) [1] aims to establish a world-class, nationwide, integrated infrastructure to support the high-performance computational needs of the defense community for research, development, testing, and evaluation. The HPCMP has established four Major Shared Resource Centers (MSRCs) to coordinate and lead this effort: ARL, CEWES, NAVO and ASC.

The Defense Department's high performance computing facilities will be supporting roughly 4,000 scientists and engineers at over 100 defense laboratories, test centers, universities, and industrial sites across the nation. In addition, the MSRCs are establishing collaborative partnerships with several civilian high-performance computing centers in order to draw needed civilian expertise into the DoD. The application areas which will be pursued by this huge community is very broad. It includes:

- Computational Structure Mechanics
- Computational Fluid Mechanics
- Computational Chemistry and Materials Science

*This work was supported in part by the European Research Office of the United States Army, Contract No. WK2Q6C-8076-EE01 "Optimal Program Scheduling on Parallel and Distributed Systems" and carried out by Oxford Parallel in collaboration with the Major Shared Resource Center, United States Army Research Laboratory, Aberdeen, Maryland.

- Computational Electromagnetics and Acoustics
- Climate/Weather/Ocean Modelling
- Signal/Image Processing
- Forces Modelling and Simulation/C4I
- Environmental Quality Modelling and Simulation
- Computational Electronics and Nanoelectronics
- Integrated Modelling and Test Environments

The MSRCs will need to solve a number of major technical and systems management challenges in order to achieve the ambitious goals of this program. In particular, they will need to adopt or develop new methodologies, tools and techniques for the following problems:

- Benchmarking and procurement of scalable parallel computing systems.
- Automatically allocating parallel codes to the most appropriate parallel machine for that code, i.e. the one which has the highest acceptable computation to communication ratio. This will in general be the cheapest machine. In the future, this will be a critical component in achieving cost-effective utilisation of major shared parallel resources in a large user community such as the DoD or the DoE community. Running a compute bound code on an expensive T3E or running a communication bound code on a cheap network of workstations are two examples of what will need to be avoided by any management responsible for the efficient use of such scalable computing resources.
- Optimised job scheduling and resource allocation for parallel codes running on shared parallel architectures.

The scientists and engineers developing application programs will need new tools, courses and training materials which will enable them to develop portable parallel programs which can achieve high performance, in a predictable way, on any of the parallel systems available at the MSRCs.

In 1996, Oxford Parallel and the United States Army Research Laboratory Major Shared Resource Center agreed to collaborate on a small initial project to explore the possible role of BSP computing in the areas of high

performance computing covered by the HPCMP and, in particular, on the prospects for optimal program scheduling on parallel and distributed systems through the automatic derivation and use of BSP cost signatures.

This report briefly describes the results of the project and refers to the various documents produced.

2 Benchmarking

A major challenge for the MSRCs will be to develop a rational framework for the benchmarking and procurement of scalable parallel computing systems. The BSP model [2, 9] provides a breakthrough in this area. It greatly simplifies the task of evaluating the performance of parallel architectures, and provides those responsible for procurement with a simple, objective and rational basis for their decisions. In [6] we show how the three key parameters of a parallel architecture – node performance, bandwidth and latency are captured by the BSP parameters s, g and l respectively. A detailed table giving the BSP parameters for many current machines is also provided.

3 Signatures and Scheduling

In [8] we describe how BSP and MPI cost signatures can be automatically produced using profiling tools. We also discuss how these signatures can be used to automatically allocate parallel codes to the most appropriate parallel machine for that code, i.e. the one which has the highest acceptable computation to communication ratio. Finally, we discuss the future prospects for using more complex cost signatures to optimise job scheduling and resource allocation for parallel codes running on shared parallel architectures.

4 Scalable Parallel Programming

In [5] we describe BSPlib, a new industry standard for scalable parallel programming. The use of BSPlib in the development of parallel applications will enable the scientists and engineers in the HPCMP to produce scalable programs which will run unchanged, with optimal performance, on any of the parallel machines at the MSRCs.

In [4] we describe the main characteristics of BSP programming. We also compare the BSP approach with the two other main approaches to parallel programming – data parallelism (HPF) and message passing (MPI).

5 Training and Technology Transfer

In the week beginning 15th July 1996, Oxford Parallel hosted a week long visit by Dr J P Collins of Army Research Laboratory (ARL), Aberdeen, MD, Dr R L Post of ARL and HPTI Inc, and Dr M Behr of the Army High Performance Computing Centre in Minnesota. During the week, a large number of members of Oxford Parallel described their work on BSP computing, on computational fluid dynamics, and on computational electromagnetics. They also described how the BSP approach is being used to develop portable and scalable parallel software systems for industrial applications. Documentation on these various areas was also provided. Some preliminary work on analysing the BSP cost of a major ARL application code was carried out. There was also some preliminary discussions between the visitors and the members of Oxford Parallel on the approach to be taken in the work on job scheduling for parallel and distributed systems

On Thursday 10th October 1996, Professor W F McColl and Dr J Hill of Oxford Parallel gave a one-day course at Aberdeen, MD on "Scalable Computing using the Bulk Synchronous Parallel (BSP) Model". The course participants (approximately 25 people, including project managers, researchers and programmers) were mainly drawn from the Army Research Laboratory and other Department of Defense Laboratories. Each participant was provided with a number of papers and other documentation produced by the course lecturers and their colleagues at Oxford Parallel. Local organisation for the course was provided by the Army Research Laboratory. The course covered the following topics:

1. Models and their role in parallel computing. Parallel architectures and their convergence. The BSP model: supersteps, cost models. Advantage of supersteps for correctness and performance.

2. BSP programming styles. Comparisons with message passing and data parallelism. BSPlib - The BSP Worldwide Standard Library.

3. Structured communications: broadcasting, reduction, prefix computations. BSP scheduling of parallel computations: solution of triangular systems, matrix multiplication, LU decomposition, algebraic path problem.

4. Implementation of BSPlib on distributed memory and shared memory architectures. The BSP Programming Environment: tools for profiling, benchmarking and debugging.

5. Design and analysis of efficient BSP algorithms: optimal matrix multiplication, matrix-vector multiplication (dense, sparse),FFT, sorting.

6. Use of BSP software tools in developing high performance applications.

On Friday 11th October, Professor McColl and Dr Hill gave a half-day practical workshop on BSP programming at ARL. Both the course and the workshop proved to be very successful and resulted in a significant increase in awareness of the potential of the BSP approach for the development of portable and scalable parallel software. The slides produced for the course have been reproduced as a project report [7].

Professor McColl lectured on "The BSP Approach to Scalable Parallel Programming" at the Department of Defense HPCMP Users Meeting hosted by the Aeronautical Systems Center MSRC at the national Center for Supercomputing Applications in Illinois on 6-8 November 1996.

A comprehensive World Wide Web site has been set up by Oxford Parallel which shows the BSP resources available to the DoD community. The site can be accessed via the WWW pages of the Army Research laboratory MSRC. The site contains links to tutorial material, technical papers, software tools and libraries. There are also pointers to various other sources of additional information on BSP programming. The contents of the site has been documented in [3]. One of the pointers is to the work of the Oxford BSP Group, which is the world's largest research group in the area of BSP computing. It is active in a number of areas related to the interests of the DoD community. These include:

- Unified scalable parallel programming environments.
- Design, analysis and implementation of BSP algorithms.
- PRAM simulation and automatic memory management in BSP computing.
- Scalable I/O for BSP programming.
- Formal methods for the specification and design of BSP programs.
- Parallel discrete event simulation and applications.
- Parallel object database systems.
- Scalable parallel linear algebra computations.

- Quantitative comparison of BSP with message passing approaches (e.g. PVM, MPI).
- BSP computers with multiple workloads.
- Parallel computational fluid dynamics.

References

[1] A K Jones. Modernizing high-performance computing for the military. *IEEE Computational Science and Engineering*, pages 71–74, Fall 1996.

[2] W F McColl. Scalable computing. In J van Leeuwen, editor, *Computer Science Today: Recent Trends and Developments. LNCS Volume 1000*, pages 46–61. Springer-Verlag, 1995.

[3] W F McColl. BSP–a new industry standard for scalable parallel computing: WWW resources. OPSPDS Technical Report 6-97, Oxford Parallel - US Army Research Laboratory, October 1997.

[4] W F McColl. Parallel programming paradigms: BSP, message passing and data parallelism. OPSPDS Technical Report 4-97, Oxford Parallel - US Army Research Laboratory, August 1997.

[5] W F McColl. Scalable, portable and predictable parallel programming with BSPlib. OPSPDS Technical Report 1-97, Oxford Parallel - US Army Research Laboratory, May 1997.

[6] W F McColl and J M D Hill. BSP and its role in the benchmarking and procurement of scalable computing systems. OPSPDS Technical Report 3-97, Oxford Parallel - US Army Research Laboratory, August 1997.

[7] W F McColl and J M D Hill. Scalable parallel programming using the BSP model: A tutorial introduction. OPSPDS Technical Report 5-97, Oxford Parallel - US Army Research Laboratory, October 1997.

[8] W F McColl and J M D Hill. Signatures and the optimal scheduling of BSP and MPI programs. OPSPDS Technical Report 2-97, Oxford Parallel - US Army Research Laboratory, June 1997.

[9] L G Valiant. A bridging model for parallel computation. *Communications of the ACM*, 33(8):103–111, 1990.